近红外光谱分析在玉米单籽粒品种真实性鉴定中的影响因素

赵怡锟[1]，于燕波[1]，申兵辉[2]，杨勇琴[1]，艾俊民[1]，严衍禄[3]，康定明[1*]

1. 中国农业大学农学院，北京 100094
2. 中国农业大学理学院，北京 100094
3. 中国农业大学信息与电气工程学院，北京 100094

摘 要 以不同存储时间，同一产地及收获时间的10个玉米品种种子为对象，研究存储时间在玉米单籽粒近红外光谱真实性鉴定中，对近红外光谱分析技术检测结果的影响。利用1月份光谱数据建立品种真实性鉴定模型(单月建模)，分别鉴定2到12月的相同品种，原始光谱采用平滑、一阶差分和矢量归一化进行预处理，PLS-DA建立模型进行分析比较，结果显示，正确鉴定率均呈逐月下降的趋势，同一品种的同一种子批，由储藏开始建立的品种真实性鉴定模型已无法对储藏11个月后的该种子批进行高准确度的鉴定，储存时间由1个月增加至11个月时，模型的平均正确鉴别率降低26.27%，这说明玉米种子的存储时间越长将降低应用近红外光谱鉴定品种真实性的鉴定准确度。另外，本研究发现玉米种子存储时间越长，导致同一品种种子样品的光谱数据在空间分布上产生差异，光谱数据离散化更明显，重复性一致性越低，使得玉米种子的真实性鉴定结果的准确性越低。通过扩充建模集中易受干扰的信息的范围，即将1年内在不同时间段里随取样时间变化而导致的在不同环境因素、仪器因素及种子样品等变化因素下采集到的光谱数据，均扩充到建模光谱数据中，以增加根据扩充数据建立的近红外光谱预测模型的包容性。通过1月与2月建模集联合后建立的包容性模型(联合建模)，之后分别对2016年3月—12月测试集的样品进行鉴定，之后逐月增加建模集光谱数据，并对非建模集所在月份进行逐月鉴定，以京科968为例，结果表明，模型对建模集相邻月份的测试准确度较高，之后逐月降低。当建模集内加入1到6月份建模集内的特征光谱后，包容性模型的平均正确鉴别率可稳定在92%以上。通过以上方式，对10个玉米品种进行了测试，结果表明，包容性模型对于玉米种子真实性的正确识定率相较于单月模型均有明显提高。J92与XY211的平均正确鉴别率分别提高11.58%与7.71%。将2016年整年的光谱数据均加入包容性模型的建模集中，使测试集玉米杂交种2017年的平均正确鉴别率达到94.68%，自交系达到95.03%，为进一步研发专用模型和实用设备提供基础。

关键词 近红外；玉米；单籽粒；模型稳定性；真实性

中图分类号：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)07-2229-06

引 言

作物品种真实性的鉴定，传统上多用表型鉴定和当下广为应用的DNA或蛋白质分子鉴定。然而，这两种方法不仅耗时费力，对环境也有污染[1-3]。而近红外光谱检测具有快速、无损样品、环境友好的特点，已应用于多种作物种子的品种真实性鉴定，以及种子营养成分含量的检测[4-6]。杨传得等利用近红外反射光谱检测花生种子含水量，模型决定系数(R^2)为0.936 2[7]。王传梁等用近红外漫反射光谱分析大米籽粒的脂肪含量，决定系数为0.988 9[8]。张初等基于反向传播神经网络进行了近红外光谱对于西瓜种子品种真实性的鉴定[9]。但近红外光谱检测因易受过程处理和环境的影响，光谱信号波动大，检测结果的可重复性和稳定性较差，在归类判别类型的定性检测中应用受到极大限制。同一种子批随着时间的推移，其光谱会产生变化，因此某一时间建立的模型对其他时间所采集的光谱的鉴别率会下降，最终导致模型的稳定性下降；但目前我国对近红外光谱分析模型稳定性的研究不多，影响该项技术的实际应用。

玉米是主要粮食作物之一，玉米种子真实性的检测鉴

收稿日期：2019-07-02，**修订日期**：2019-11-06

基金项目：国家重点研发计划[(2017YFD0102001-3)]资助

作者简介：赵怡锟，1991年生，中国农业大学农学院博士研究生 e-mail：233784675@qq.com

*通讯联系人 e-mail：kdm@cau.edu.cn

定，在良种培育、种子生产和经营中需要频繁进行。目前玉米自交系和杂交种子的真实性鉴定，种子生产和经营中常采用特征同工酶聚丙烯酰胺蛋白电泳凝胶谱带技术[10]，以及通过 DNA 特征位点的 PCR 扩增来鉴定识别品种真实性[11-12]，但上述方法技术复杂，成本较高，耗时长，不能进行实时在线分析。近红外光谱技术可以用于玉米真实性鉴定。唐兴田等利用仿生模式识别建模，进行了玉米品种真实性鉴定，平均正确识别率和平均正确拒识率达到 94.5%和 96.71%。但还不能对非建模时间段采集的样品进行预测，对不同存储时间下玉米种子的预测也没有研究，不同种子存储时间带来的模型不稳定性问题没有解决。

为此，我们将近红外光谱检测技术用于玉米品种真实性鉴定，选用 10 个相同地点和相同时间生产的玉米杂交种及自交系的种子，研究玉米种子存储时间的长短对近红外光谱检测结果可重复性和稳定性的影响，进而通过一年间不同时段的对相同种子批样品的持续光谱采集，最后针对采集到的所有近红外光谱数据，进行联合建模，由此来判别分析下一年度采集相同样品的光谱检测结果，用模型对检测样品的正确鉴别率来分析玉米种子存储时间对模型预测结果的稳定性和可重复性的影响。

1 实验部分

1.1 材料与光谱采集

玉米种子为收获于 2015 年产自甘肃张掖的玉米杂交种和自交系种子，共 12 个品种，1440 份样品。10 个品种作为实验组，2 个作为对照组，对照组将作为非标准样品掺入建模集，参与真实性建模，其中 4CV-C72 作为杂交品种非标样，郑 58 作为自交系非标样，详细信息见表 1。

表 1 12 个品种详细信息

Table 1 Twelve varieties of maize samples

品种名称	类别	样品数	简称	品种名称	类别	样品数	简称
京科 968	杂交种	120	JK968	6WC	自交系	120	6WC
龙单 16	杂交种	120	LD16	京 92	自交系	120	J92
玉源 5	杂交种	120	YY5	京 724	自交系	120	J724
丰禾 10	杂交种	120	FH10	4CV	自交系	120	4CV
信玉 211	杂交种	120	XY211	昌 7-2	自交系	120	C72
4CV-C72(对照组)	杂交种	120	4CV-C72	郑 58(对照组)	自交系	120	Z58

近红外光谱数据采集采用 MicroNIR 1700 近红外微型光谱仪，数据分析软件为 Matlab®(Mathworks，Inc，U. S.)。

所有收获的样品在室内通过均匀干燥，再加湿的方法，使样品水分处于 10%～11%，密封后，平衡水分 1 周。样品筛选，每个品种挑选籽粒完整，颜色一致的种子 120 粒。光谱采集前，光谱仪需预热 45 min，之后进行单籽粒光谱数据采集。采集方式为漫反射，积分时间为 10 000 μs，积分次数为 400 次，采集时间 4 秒，在每粒种子胚面位置采集光谱 3 次，取平均值，每组种子采集完光谱，单粒种子逐一做标记，室温通风处存放。

为使 1 年中采集的光谱数据时间相隔均匀，以月为单位，1 年分为 12 个采集光谱时间段，每月采集 4 次，分别在每月 4，11，18 和 25 日上午(9 点—11 点)，下午(14 点—16 点)，晚上(19 点—21 点)3 个时间点进行，以此作为每月的光谱数据。每天的单个时间点每个品种采集 10 条光谱，1 天 3 个时间点，每天合计 30 条光谱，每月 120 条光谱。

1.2 方法

对原始光谱进行平滑预处理(平滑窗口长度为 9)，消除噪声干扰，然后使用一阶差分导数(差分宽度为 9)放大差别，消除基线和其他背景干扰，分辨重叠峰，提高分辨率和灵敏度，最后用矢量归一化降低同一品种玉米籽粒光谱若干次采集间的差别。

在上述预处理的基础上，采用 PLS-DA 建立品种真实性鉴别模型，以月为单位，从每品种单月采集的 120 条光谱中选取前 90 条作为建模集，用于建立预测模型，后 30 条作为测试集。即 JK968 的建模集由 2016 年 1 月采集的 90 条 JK968 与 90 条 4CV-C72 光谱组成，测试集由 2 月采集的 30 条 JK968 与 30 条 4CV-C72 光谱组成，以此类推，其他各品种建模集和预测集相同方法建立。

本研究用正确鉴别率(correct identification rate，CIR)来评价模型性能，其计算方式为：(模型正确鉴别样品个数＋模型正确拒识样品个数)/参与测试样品总数。最终通过主成分分析以及模型正确鉴别率来综合分析玉米种子存储时间对近红外光谱分析结果的影响程度。

2 结果与讨论

2.1 不同玉米种子存储时间的种子真实性鉴定结果

为了探究玉米籽粒存储时间对玉米品种真实性的近红外光谱单籽粒预测模型稳定性的影响，利用 1 月份光谱数据建立品种真实性鉴定模型(单月建模)，分别鉴定 2 到 12 月的相同品种，结果如图 1 所示，纵坐标表示 1 月份建立的单月模型对 2 到 12 月每月测试集鉴定的正确鉴别率，结果显示，正确鉴定率均呈逐月下降的趋势，其中，J724 的正确鉴别率由 98.33%下降至 63.33%，下降较明显。同时，同一品种的同一种子批，由储藏开始建立的品种真实性鉴定模型已无法对储藏 11 个月后的该种子批进行高准确度的鉴定(所有样品 1 月份单月模型对 12 月测试集鉴定的正确鉴别率均在

83.33%以下)。这说明玉米种子的存储时间越长将降低应用近红外光谱鉴定品种真实性的鉴定准确度。

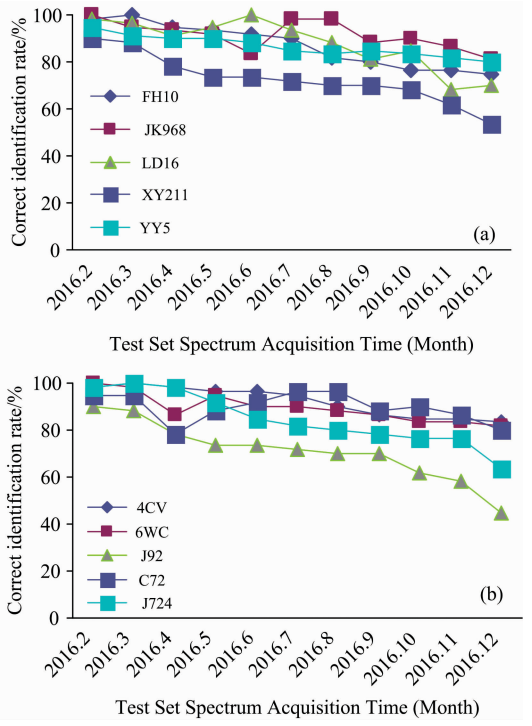

图 1　基于 1 月份单月所建模型对 2 月至 12 月测试集的鉴别结果

(a)：杂交种；(b)：自交系

Fig. 1　Identification results of model based on data in Jan，the tested data collected in Feb to Dec

(a)：Hybrids；(b)：Inbred lines

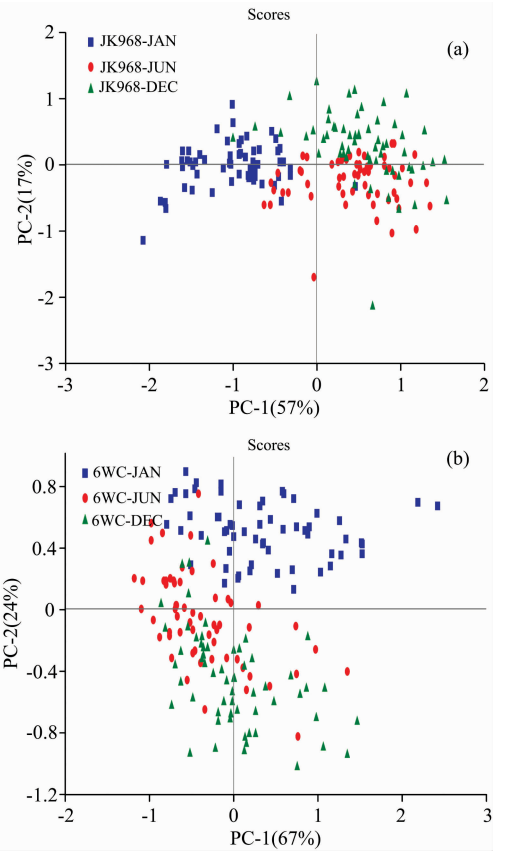

图 2　1 月，6 月，12 月预处理后光谱数据 PCA 分析结果

(a)：JK968；(b)：6WC

Fig. 2　PCA scores plot for pretreatment data collected in Jan，Jun and Dec

(a)：JK968；(b)：6WC

2.2　玉米种子存储时间变化导致光谱数据离散

为研究模型随玉米种子样品存储时间的变长，鉴定结果不稳定的原因，用 10 个品种的 2016 年 1 月到 12 月的光谱数据来剖析，随着玉米种子储存时间的推移，样品光谱数据的变化。

以杂交种 JK968，自交系 6WC 为例，图 2 为 JK968 与 6WC 的原始光谱经过预处理后，通过主成分分析将光谱数据降维后得到的 2 维图(图中“□”表示 1 月份光谱数据，“○”表示 6 月份光谱数据，“△”表示 12 月份光谱数据)，由图可见，不管是 JK968 或是 6WC，玉米种子不同存储时间样品的光谱呈现不同程度的离散，1 月份和 6 月份 JK968 的光谱数据在特征空间中有部分重叠，而 12 月份的光谱数据与 1 月份的相对距离较远，几乎在空间内分开，说明无论是杂交种还是自交系，应用近红外光谱进行玉米种子真实性单籽粒鉴定时，同一品种样品的光谱数据会随着玉米种子样品的存储时间推移产生离散变化，在玉米种子样品存储 11 个月后，这种离散变化更加显著，重复性一致性越低，使得玉米种子的真实性鉴定结果的准确性越低。

2.3　扩充模型包容性，增加鉴定结果稳定性

同一种子批在 1 年的不同时段月份采集的光谱数据产生变化，即延长玉米种子的储存时间会导致模型的正确鉴别率下降。针对上述问题，用扩充模型的包容性来消除因建模数据与检测样品数据时间跨度较大对近红外光谱检测玉米种子品种真实性鉴定的影响。用稳定性(也称容变性)来评价玉米种子品种真实性鉴定的近红外光谱模型，模型的稳定性指模型适配范围的大小，主要决定于建模光谱集中不确定信息(背景信息)变动的范围。

背景信息是检测样品存储时间变量所带来的，随着玉米种子存储时间的延长，存储环境变化的温度、湿度，空气中水分等外界因素都可能会引发玉米种子内部有机物含量的变化，而由此干扰不同时间段采集到的近红外光谱的一致性，从而使得在这些光谱数据基础上建立的模型的稳定性。我们通过扩充建模集中受干扰的信息的范围，即将 1 年内在不同时间段里，也即在不同环境因素、仪器因素及种子样品等变化因素下采集到的光谱数据，均扩充到建模光谱数据中，以增加根据扩充数据建立的近红外光谱预测模型的包容性。

包容性模型的具体建立方法，以 JK968 为例，如表 2 所示，模型编号 1-2 表示 1 月与 2 月建模集联合后建立的包容性模型(联合建模)，之后分别对 2016 年 3 月—12 月测试集的样品进行鉴定，之后逐月增加建模集光谱数据，并对非建模集所在月份进行逐月鉴定，例如，1—11 表示将 1 月至 11

月的建模集进行联合建模，对 12 月测试集进行鉴定。可以看出，模型 1-2 至 1-5 整体呈现出图 4 相同的趋势，即模型对建模集相邻月份的测试准确度较高，之后逐月降低。当建模集内加入 1 月—6 月份建模集内的特征光谱后，包容性模型的平均正确鉴别率可稳定在 92%以上。

表 2 联合建模的真实性检测结果(JK968)

Table 2 Identification results of cross-modeling based on JK968 (Unit: %)

编号	日期										
	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	平均值
1-2	98.75	100	98.75	92.5	91.25	87.5	83.75	78.75	77.5	71.25	88
1-3		93.75	90	83.75	96.25	96.25	83.75	73.75	72.5	76.25	85.14
1-4			96.25	93.75	93.75	90	86.25	88.75	81.25	81.25	88.9
1-5				95	96.25	93.75	91.25	87.5	83.75	78.75	89.46
1-6					100	97.5	95	90	93.75	90	94.38
1-7						100	98.75	83.75	91.25	86.25	92
1-8							98.75	96.25	92.5	92.5	95
1-9								98.75	100	96.25	98.33
1-10									100	97.5	98.75
1-11										98.75	98.75

通过以上方式，对 10 个玉米品种进行了测试，结果见表 3，ACIR(单月建模)表示 1 月单独建模分别鉴定 2 月—12 月测试集的正确鉴别率平均值，ACIR(联合建模)表示所有包容性模型对其非建模所在月份测试集鉴定后正确鉴别率的平均值，可以看出，包容性模型的平均正确鉴别率相较于单月模型均有明显提高，其中 J92 与 XY211 的平均正确鉴别率分别提高 11.58%与 7.71%。

表 3 联合建模的平均正确鉴别率

Table 3 Average identification results of cross-modeling (Unit: %)

品种名称	ACIR (单月建模)	ACIR (联合建模)	品种名称	ACIR (单月建模)	ACIR (联合建模)
FH10	91.14	91.88	4CV	91.25	92.34
JK968	88.07	92.87	6WC	89.32	92.32
LD16	88.05	89.88	C72	90.11	91.12
XY211	72.5	80.21	J92	71.14	82.72
YY5	86.25	87.02	J724	84.66	89.56

表 4 基于 2016 整年联合建模的模型对于 2017 年测试集鉴别结果

Table 4 Identification results of cross-modeling by dataset of 2016 whole year, data collected in 2017 for test (Unit: %)

样品名称	平均正确鉴别率	样品名称	平均正确鉴别率
FH10	96.99	4CV	90.50
JK968	92.86	6WC	93.02
LD16	89.54	C72	97.17
XY211	96.30	J92	97.63
YY5	97.72	J724	96.82
平均值	94.68	平均值	95.03

总的看来，为使近红外光谱的玉米种子真实性单粒鉴定的正确鉴别率进一步提高，将 2016 年整年采集的光谱数据均加入到包容性模型的建模集中，并对模型进行为期一整年的测试，测试集为 2017 年 1 月—12 月采集的光谱，测试结果见表 4，表中所示平均正确鉴别率为 2016 年包容性模型对样品 2017 年每个月测试集鉴定后的正确鉴别率平均值。可以看出，除了 LD16 的平均正确鉴别率为 89.54%，其他样品均在 90%以上，另外，玉米杂交种全年的平均正确鉴别率为 94.68%，自交系为 95.03%，10 个玉米品种的平均正确鉴别率为 94.86%。

3 结 论

以同一产地，同一收获时间的 10 个玉米品种为对象，研究玉米种子的存储时间对近红外光谱玉米种子品种真实性单籽粒鉴定的影响程度。结果表明：种子存储时间的延长会直接降低近红外光谱对玉米种子品种真实性单籽粒鉴定的正确鉴别率。故此，提出了扩充模型的包容性，来解决玉米种子存储时间变量对模型稳定性的影响。通过联合建模，包容性模型较于单月模型的正确鉴别率明显提高。同时，为进一步提高模型的正确鉴别率，将 2016 年整年的光谱数据均加入包容性模型的建模集中，提高模型应对光谱采集时间、环境等可能不一致条件的应变能力，增强模型的稳定性，使测试集玉米杂交种 2017 年的平均正确鉴别率达到 94.68%，自交系达到 95.03%，为进一步研发专用模型和实用设备提供基础。

对于玉米种子存储时间的延长所导致的光谱变化的具体原因，将在后续研究结果中报道，以期对进一步提高近红外光谱玉米种子单籽粒品种真实性鉴定模型的正确鉴别率有指导意义。

References

[1] Carew M E, Nichols S J, Batovska J, et al. Marine and Freshwater Research, 2017, 68(10): 1788.
[2] Mantelatto F L, Terossi M, Negri M, et al. Mitochondrial DNA Part A, 2018, 29(5): 805.
[3] Pan Y B. Agronomy, 2016, 6: 28. doi: 10.3390/agronomy6020028.
[4] Malegori C, Buratti S, Benedetti S, et al. Talanta, 2020, 206: 120208.
[5] Zhang H, Duan Z, Li Y Y, et al. Royal Society Open Science, 2019, 6(10): 191132.
[6] ZHOU Guang-hua, ZHU Da-zhou, WANG Cheng(周光华, 朱大洲, 王 成). Journal of Anhui Agricultural Sciences(安徽农业科学), 2010, 38(28): 15475.
[7] YANG Chuan-de, YU Hong-tao, GUAN Shu-yan, et al(杨传得, 于洪涛, 关淑艳, 等). Journal of Peanut Science(花生学报), 2012, 41(1): 6.
[8] WANG Chuan-liang, CHEN Kun-jie(王传梁, 陈坤杰). Cereals and Oils Processing(粮油加工), 2007, (2): 62.
[9] ZHANG Chu, LIU Fei, KONG Wen-wen, et al(张 初, 刘 飞, 孔汶汶, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2013, 29(20): 270.
[10] TENG Tian-yong(滕天涌). Seed World(种子世界), 2003, (4): 22.
[11] YU Xin-yan, WANG Feng-ge, ZHAO Jiu-ran, et al(于新艳, 王凤格, 赵久然, 等). Molecular Plant Breeding(分子植物育种), 2007, 5(3): 443.
[12] WANG Feng-ge, ZHAO Jiu-ran, WANG Lu, et al(王凤格, 赵久然, 王 璐, 等). Journal of Agricultural Biotechnology(农业生物技术学报), 2007, 15(6): 964.

Influence Factors in Near-Infrared Spectrum Analysis for the Authenticity Identification of Maize Single-Kernel Varieties

ZHAO Yi-kun[1], YU Yan-bo[1], SHEN Bing-hui[2], YANG Yong-qin[1], AI Jun-min[1], YAN Yan-lu[3], KANG Ding-ming[1]*

1. College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China
2. College of Science, China Agricultural University, Beijing 100094, China
3. College of Information and Electrical Engineering, China Agricultural University, Beijing 100094, China

Abstract The study, targeting at 10 Maize varieties with different storage time and the same origin and harvest time, aims to study the effects of storage time on the results of the near infrared spectrum analysis technology applied in the near-infrared spectrum authenticity identification of maize single-kernel varieties. The authenticity model (monthly modeling) of breeds was established by using spectral data from January to identify the same samples which spectral data from February to December. The original spectrum was pre-processed by smoothing, first order difference and vector normalization. PLS-DA was used to establish the model for analysis and comparison, the results showed that the correct identification rate was decreasing month by month. The average correct identification rate of the model is reduced by 26.27% when the storage time is increasing from 1 month to 11 months, Which indicates that the longer the storage time of maize seeds is, the lower the accuracy of the near-infrared spectrum authenticity identification will be. This research also indicated that with the increase of the storage time of maize seeds, the spatial distribution of the spectral data of the same species but at different storage time is different. The discretization of spectral data becomes obvious, and the repeatability and consistency are reduced, which makes the accuracy of authenticity identification results of maize seeds is reduced. We endeavor to expand the models to centralize the range of the information that is easily interfered, that is, expand the spectral data collected under different environmental factors, instrumental factors and seed samples in different periods of time in 1 year to the modeling spectrum data to increase the inclusiveness of the prediction model of the near infrared spectrum based on the expanded data. Then, the inclusive model (joint modeling) has established by jointing the January and February modeling sets, after that, identifies the test set samples from March to December respectively, and then increases the model set spectrum data month by month, and the identifies the months that non-modeling set is located month by month. It taking JK968 as an example, the results showed that the accuracy of the model for the adjacent months of the modeling set is high, and then decreases month by month. When the feature spectrum of the model set is added from January to June, the average correct identification rate of the inclusive model can be more than 92%. In the above way, 10 maize varieties were tested, which can be seen that the correct identification rate of the inclusive model for maize seed authenticity is significantly higher than

that of the single month model. The average correct identification rate of J92 and XY211 is increased by 11.58% and 7.71%, respectively. At the same time, in order to further improve the correct identification rate of the model, this study added the spectral data of the year 2016 to the modeling concentration of the inclusive model, so that the average accuracy identification rate of maize hybrids in 2017 reached 94.68%, and the inbred line reached 95.03%, providing the basis for further developing special models and practical equipment.

Keywords Near infrared; Maize; Single-kernel; Model stability; Authenticity

(Received Jul. 2, 2019; accepted Nov. 6, 2019)

* Corresponding author